Dimension Reduction Regression Techniques for High Dimensional Data

W. Holmes Finch **Maria E. Hernández Finch** **Lauren E. Moss**
Ball State University

In some situations, researchers are faced with a regression analysis involving a large number of independent variables (predictors) relative to the sample size. Such cases are referred to as high dimensional data and can present problems for standard regression analyses, including collinearity among the independent variables, and in the extreme case an inability to obtain parameter estimates. One data analysis strategy that has been recommended for such situations is principal components analysis, which combines the predictors into a smaller number of linear combinations that are then used as independent variables themselves in a regression analysis with the original dependent measure. Recently, several variations to this approach for data reduction have been recommended for use. The goal of this paper was to describe several of these and to innovate their use with high dimensional social science datasets.

High dimensional data refers to the situation in which the number of variables to be used in an analysis approaches or exceeds the number of observations upon which measurements are made (Samet, 2006). This situation is written symbolically as $p >> n$. The consequences of trying to apply familiar tools, such as linear models, to data can include overfitting of the model to the sample, a high variance in parameter estimates, and unstable parameter estimates (Hastie, Tibshirani, & Friedman, 2011). Overfitting of the data to the sample means that results obtained for a single sample may not generalize well to the population from which the sample was drawn because the estimates are too closely tied to idiosyncrasies specific to the individual sample. When $n > p$, coefficients from linear models can be taken to generalize to the population, assuming that the sample is representative of the population, which might be the case through random selection of the sample. However, when $p >> n$, overfitting will likely occur even when the sample has been randomly selected. A second major problem in the high dimensional case is that parameter estimates will tend to have high variability (Bühlmann & van de Geer, 2011), which will manifest itself in the form of large standard errors and consequent high Type II error rates as well as estimates that are far from the actual population values. In practice, researchers might see this situation in the form of values that appear counter-intuitive (e.g., a positive regression slope when theory would suggest a negative slope), or in extreme values. Furthermore, in the context of regression, if the number of independent variables exceeds the sample size, parameter estimates may not be obtainable and standard errors will certainly not be available.

While high dimensional data has traditionally been most common in fields such as genomics (Datta, 2001), it is also possible in the social sciences for researchers to be faced with a situation in which they have many measured variables on a small number of subjects. For example, there are certain populations of individuals that are very small and/or hard to sample. In the current study, we focus on children of migrant workers. Given their high degree of mobility, and in some cases a desire to avoid contact with institutions due to fears of deportation (United States Department of Justice & Department of Education, Civil Rights Division, Office
for Civil Rights & Office of the General Counsel, 2011), such individuals may be hard to identify and to gain permission for measurements to be made. Given the difficulty of identifying these children in the first place once a sample has been obtained, the researcher may wish to gather as much data from them as possible. Other populations in the social sciences that may be relatively difficult to sample in large numbers for a variety of reasons, thereby presenting the researcher with high dimensional data, are those with certain low incidence developmental disabilities such as Autism or Rhett's syndrome. Researchers working in very sensitive areas, such as sexual behavior; substance abuse; or suicidology, may also find it difficult to obtain large samples of individuals, even while they are able to record a large number of variables for those individuals that they do sample (Solomon, Hill, Janssen, Sanders, & Heiman, 2012).

Given the possibility for researchers in the social sciences, as well as other disciplines, to be faced with the p>>n problem, the current study focuses on the demonstration of several methods for data analysis in the presence of such high dimensional data, specifically in the context of regression. When researchers want to assess the relationships of multiple independent variables with a continuous dependent variable, they frequently choose multiple regression. However, in the presence of high dimensional data, regression will degenerate as noted above. Therefore, researchers in such situations need alternative approaches for data analysis that will allow them to address their research questions. Our goal is to take an example of high dimensional data and present three such methods that are based in dimensionality reduction, including two variants of principal components analysis and partial least squares. Each of these alternatives is described in some detail below, after which each of them is applied to an actual high dimensionality data problem. We conclude the manuscript with a discussion of the results, a comparison of the outcomes provided by each method, and advice for researchers faced with high dimensional social science data.

When a researcher needs to develop a model relating multiple independent (*x*) variables to a single continuous dependent (*y*) variable, she or he may select multiple regression in order to obtain estimates of the relationships between each *x* and *y*. The standard linear regression model takes the following form in the population:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_j x_{ji} + \varepsilon_i \tag{1}$$

Here, the outcome variable,y_i, is expressed as a function of an intercept (β_0), and a set of *J* independent variables ($x_{1i} \ldots x_{ji}$), and their coefficients ($\beta_1 \ldots \beta_j$) as well as random noise (ε_i). In order to obtain sample estimates of the β_0 and β_j values, ordinary least squares (OLS) is typically used. OLS finds the model parameter estimates that minimize the function

$$E^2 = \sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji}))^2 \tag{2}$$

which can also be expressed as

$$E^2 = (y_i - \hat{y}_i)^2 \tag{3}$$

where $\hat{y}_i$ is the model predicted value of y_i.

Data Reduction

As noted above, when the sample size is small relative to the number of independent variables, a number of problems can result for OLS regression. One general approach that has proven useful for high dimensional data situations in research contexts, such as genomics (Datta, 2001) and computer science (Rosipal & Trejo, 2001), are data reduction techniques. These methods share the common trait of taking the *J* independent variables and reducing them to *m* linear combinations, where $m < J$. These linear combinations then take the role of independent variables in regression analyses in which the original *y* continues to serve as the dependent variable. Data reduction methods differ in terms of how they create the linear combinations of the *x* variables. One of the most popular data reduction methods and; thus, one that will be examined here is principal components regression (PCR), where standard principal components analysis (PCA) is applied to the set of predictors and the resultant principal components are then used as predictors in the regression analysis. An alternative to PCR, supervised PCR (SPCR), is similar in approach, but takes into account relationships of the original *x*s with *y* when creating the linear combinations. A third approach, partial least squares regression (PLSR), shares several traits with PCR and SPCR, but differs in terms of how the linear combinations are created. Each of these methods will be described in some detail below. Following these general descriptions of the methods, each will be applied to a dataset in order to demonstrate its use and the type of output that a social science researcher making use of it can expect to obtain. Finally, a discussion of the relative merits of each method will be given and recommendations for researchers faced with high dimensional data will be provided.

PCR

Perhaps the most popular and certainly one of the oldest methods of data reduction in the context of regression is PCR (Draper & Smith, 1981). It is based on PCA (Hotelling, 1933) and has been used in a variety of research scenarios including medicine (Tan, Shi, Tong, & Wang, 2005), engineering (Ongel, Kohler, & Harvey, 2008), economics (Gupta & Kabundi, 2010), and the social sciences (Weiss, Gale, Batty, & Deary, 2013). Prior to using PCA, the independent variables (x_j) must first be standardized, yielding z_j with a mean of 0 and standard deviation of 1. This standardization is necessary because PCA is not scale invariant, so that variables on a relatively larger scale, and that consequently have larger variances, will tend to dominate the resultant principal components. By placing all of the variables on a common scale, standardization eliminates this problem. After standardization, PCA is applies to the *z* variables in order to reduce the *J* independent variables into m principal components, which we will refer to as *w*. The resulting principal components can be expressed as

$$w_m = \gamma_1 z_1 + \gamma_2 z_2 + \cdots \gamma_j z_j \qquad (4)$$

where $\gamma_j =$ PCA coefficient for z_j and w_m is the *m*th principal component. There are as many principal components in a dataset as there are independent variables, and the first component will extract the largest share of variance in the *z* variables possible. The second principal component is orthogonal to the first, and extracts the largest share of the variance remaining after the first component has been identified. Subsequent principal components are estimated in the same manner being orthogonal to the previous components and extracting maximal remaining variance. The variance accounted for by the full set of principal components is equal to the total variance present in the original set of variables.

Although there are as many principal components as variables in the analysis, in practice the number, *m*, of components that are actually retained for use in PCR is typically much smaller than *J*, so as to avoid the aforementioned problems associated with high dimensional data. Selection of the optimal number of components may be done using a variety of criteria, none of which has proven to be optimal in all cases. However, based on simulation research, one of the more promising methods is parallel analysis (Horn, 1965), which was used in the current study. With parallel analysis, the eigenvalues from the PCA of the original data, which can be interpreted as measures of the variance explained by each component, are compared with eigenvalues obtained through random data that has the same marginal distributions of the variables as found in the original dataset. PCA is applied to this random data and the resultant eigenvalues for each principal component are retained. This creation of random datasets with marginal characteristics equivalent to that of the original data, but with negligible correlations among the variables, is repeated a large number of times (e.g., 1000), so as to create a distribution of eigenvalues for each principal component. Once the distribution is created, the eigenvalue from the original data for the first component is compared with the distribution of eigenvalues from the random dataset, and if the original is greater than or equal to the 95th percentile of the random distribution, the researcher would determine that at least the first component should be retained. Likewise, this step is repeated for the second component, and if the eigenvalue for the original data exceeds the 95th percentile of the random distribution, then the second component would be retained as well. This process continues until the eigenvalue from the original data does not exceed the 95th percentile of the random distribution. In addition to parallel analysis, the proportion of variance accounted for by each component was also examined in the current study in order to determine the number of components to retain.

As was described above, the coefficients, γ_j, are selected so as to maximize the explained variance in the observed *x* variables in the form of m principal components. Assuming that only a subset of the principal components accounts for most of the variance in the independent variables, *m* < *J* principal components can be selected for use as independent variables in a regression equation with the original *y* as the dependent. The resulting equation would take the form

$$y_i = \alpha_0 + \alpha_{\ 1} w_{1i} + \alpha_{\ 2} w_{2i} + \ldots + \alpha_{\ m} w_{mi} + \varepsilon_i \qquad (5)$$

The model coefficients and intercept are interpreted in the same manner as in the standard OLS model (1), except that they reflect the relationships between the linear combinations of the *x* variables and the dependent variable, rather than the *x*s themselves. The coefficients in (5) can be converted to regression coefficients ($\boldsymbol{\beta}^*$) for the centered and scaled versions of the original *x* variables using the following matrix equation

$$\boldsymbol{\beta}^* = \boldsymbol{V\alpha} \qquad (6)$$

where $\boldsymbol{\alpha}$ is the vector of coefficients from (5), and $\boldsymbol{V}$ is the eigenvector associated with the m retained principal components.

Supervised PCR

One of the commonly cited weaknesses of PCR is that it only accounts for the variation in the independent variables when developing the linear combination of predictors. Thus, while the resulting principal components maximize the explained variance among the independent variables, they ignore any relationships with the dependent variable. Therefore, the resulting PCR equation (5) may not optimize prediction or explanation of *y*, resulting in a regression solution that does not fully address the relationships between the outcome and the independent variables (Yu, Yu, Tresp, Kriegel, & Wu, 2006). An alternative approach that has been described in the literature is supervised principal components regression (SPCR), which is based on standard PCR as described above, but that incorporates the relationships between each x and y when developing the principal components. The primary advantage that has been noted for SPCR versus PCR is that the former incorporates information about relationships between the *x*s and *y* into the estimation of the principal components. Thus, when the regression equation for SPCR is estimated, it should yield more accurate and less noisy predictions of *y* than is true with the equation resulting from PCR.

First, as with PCR, the original x_j variables must be standardized to have a mean of 0 and a standard deviation of 1, thus, forming z_j. The SPCR algorithm then involves the following steps:

1. Estimate $y_i = \beta_0 + \beta_j z_{ji}$ (7)
   for each independent variable individually using simple linear regression.
2. Compare each β_j from (7) with a predetermined threshold, *t*.
3. Retain z_j only if β_j exceeds *t*.
4. Compute the first *m* principal components, $w_1, w_2, \ldots, w_m$, using only the retained z_j from step 3.
5. Use a jackknife cross-validation procedure to determine the optimal value of *t*.
6. When the appropriate number of principal components is fit, estimate the model

$$y_i = \alpha_0 + \alpha_1 w_{1i} + \alpha_2 w_{2i} + \ldots + \alpha_j w_{mi} \quad (8)$$

which corresponds to (5) in form, though, not in terms of actual values of w_{mi}.

Clearly, the primary difference between SPCR and PCR is that the development of the principal components in the former involves only those variables that have a demonstrated relationship with *y*, whereas for the latter procedure all of the *x* variables are involved. In order to determine *t*, the threshold for retaining predictors in the model, a jackknife cross-validation procedure is used typically (Hastie et al., 2011). This method, which is sometimes also referred to as "leave one out" cross-validation, involves fitting the model separately excluding each individual, *i*, in turn, and then obtaining a predicted value of *y* for individual *I* who has been left out of the equation. The squared difference between the fitted value, $\hat{y}_i$ and the actual value, y_i, is then calculated. This process is repeated for each member of the sample and the sum of the squared differences is then obtained:

$$E^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (9)$$

This jackknife procedure is then repeated for various values of *t* (where variables with coefficients from step 1 above less than *t* are not included in the final regression model), and the optimal setting is the one that minimizes E^2.

The model in (8) expresses the relationship between the individual principal components and the original dependent variable. However, researchers using this technique might also be interested in the relative importance of the individual *z* variables with respect to *y*. Simply because a variable had a simple linear regression coefficient in excess of the threshold, *t*, does not imply that it is truly important in terms of understanding *y* when considered with the entire set of predictors (Bair, Hastie, Paul, & Tibshirani, 2006). Bair et al. suggested as a measure of independent variable importance (imp_j), the inner product between each *z* and each principal component w_m. Variables with larger values of imp_j contribute more to explaining the variance in *y*. One point to note here is that imp_j values can be calculated for each of the *x* variables, even if they did not initially meet the threshold test in step 3 of the SPCR algorithm.

The importance score can then be used to further winnow the number of independent variables from the original set of *z*s in order to achieve the most parsimonious model possible that still accounts for the maximal amount of variation in *y*. This winnowing occurs subsequent to application of the 6 step algorithm described above. Creating such a reduced model may be of particular importance when the number of independent variables is large and the researcher is particularly interested in the individual relationships of the most important of these with the dependent variables. If the final set of predictors to be interpreted is very large, it can make understanding the relationships in the data difficult. On the other hand, if the number of predictors is reduced to a relatively small number that can be combined into principal components accounting for as much (or nearly so) variation in y as did the original solution, the subsequent interpretability of the results might be greatly enhanced. The *x* variables to retain based on *imp$_j$* are determined by comparing the coefficients to a threshold value, □, much as was done in step 3 of the algorithm using *t*. The optimal value of □ is selected using cross-validation by maximizing the log-likelihood ratio statistic for the left out observation. The difference in the cross-validation method used here and the jackknife method described earlier is that rather than removing individual observations from the data and then running the analysis, a sample of individuals from the original data set were randomly selected to be in the cross-validation sample, so as to ensure that a sufficient number of data points are available for computing the log-likelihood ratio. Furthermore, in keeping with recommendations by Bair et al. (2006), this random selection of individuals into the cross-validation sample was carried out 5 times, with the log-likelihood ratio statistic was calculated for each, and the results were averaged. In this example, 25% of the data points were randomly assigned to the cross-validation sets each time.

Once the appropriate value for □ is selected, only those variables with importance scores in excess of it are retained into the further reduced set of predictors. It should be noted that because *imp$_j$* values can be calculated for all of the *x* variables, including those not retained in step 3 of the algorithm, these variables could theoretically reappear in the reduced predictor equation even if they were not in the original. However, in keeping with recommendations by Bair et al. (2006), we will only consider those variables that were retained in step 3 for retention in the final reduced set. This practice was recommended because (1) it allows for a direct comparison of the reduced model to the original, in terms of explained variance in *y*, and (2) it frequently yields a similar solution to an unsupervised principal components analysis, thereby, negating the benefits of the supervised process (Bair et al.). The final, reduced model is then the one that researchers would typically interpret, both in terms of the relationships of individual *x* variables to y and with regard to obtaining predicted values of *y*, should that be of interest.

Partial least squares regression (PLSR)

In addition to PCR and SPCR, there exists a third approach for dimension reduction based regression, PLSR. As with PCR and SPCR, this methodology has found relatively widespread use among researchers in the natural sciences and medicine, where high dimensional data are fairly common (e.g., Carin, et al., 2012; Moon et al., 2007). It has not been as widely used in the social sciences, although it would seem to have potential application for high dimensional data problems in those fields as well. PLSR is similar in many respects to SPCR in that it derives linear combinations of the *x* variables accounting for their relationships with *y*. First, as with both PCR and SPCR, the original *x$_j$* variables must be standardized to have a mean of 0 and a standard deviation of 1, thus, forming z$_j$. Next, the following steps are followed:

1. Compute φ_{1j} for each IV, z_j. This value is obtained by finding the singular value decomposition of the cross product matrix between the independent and dependent variables,
$$\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{Y} \tag{10}$$
thereby accounting for the variances in both the *x*s and *y* and the covariances among them.
2. Compute $$w_1 = \sum_{j=1}^{J} \varphi_{1j} z_j \tag{11}$$
3. Orthogonalize all z_j with respect to w_1
4. Obtain φ_{2j} and compute $$w_2 = \sum_{j=1}^{J} \varphi_{2j} z_j \tag{12}$$
such that w_2 is orthogonal to w_1
5. Repeat steps 3 and 4 in order to obtain $w_1$to w_m, all of which are orthogonal
6. Fit the regression model $$y_i = \alpha_{\ 0} + \alpha_{\ 1} w_{1i} + \alpha_{\ 2} w_{2i} + \ldots + \alpha_{\ j} w_{mi} \tag{13}$$

PLSR, like SPCR, uses information about the variance in the independent variables as well as relationships of the independent variables to the dependent in order to create the components that are ultimately used in the regression analysis. In addition, PLSR is differentiated from SPCR in that it also accounts for the variance in *y* as well as that in *x*. This focus on maximizing explained variance in the *x*s and *y*, as well as the covariances between each *x* and *y*, makes PLSR a potentially more accurate predictive tool than PCR (Frank & Friedman, 1993) and provides a somewhat different solution than SPCR. Researchers have found that in many cases PLSR and PCR yield similarly accurate predictions of *y*, though, PLSR does so with fewer variables (Hastie et al., 2011).

In order to determine the number of linear combinations that should be retained using PLSR, jackknife cross-validation is used as described for SPCR. In this case, the number of linear combinations, *m*, to be retained is determined by assessing the root mean square prediction error

$$\text{RMSPE}=\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{n}} \tag{14}$$

for each number of linear combinations, based on the jackknife cross-validation. The optimal number is identified as the point at which adding linear combinations does not appreciably decrease RMSPE. Obviously, this is a subjective judgment made by the researcher and may differ from one researcher to another. An example of this process is presented below.

Method

Parent permission and child assent was obtained for 17 elementary age Latino/a migrant students in a STEM (science, technology, engineering, and mathematics) digital storytelling summer camp in a Midwestern state. Participants in the present study were 13 children for whom all measures were successfully collected. One potential participant from the original 17 was eliminated from the analyses due to generating a significantly below average ability score (i.e., more than 3 SDs below the measure's standardization mean). Three students were not included in analyses due to attrition as their parents had to move to find additional work before all of the data could be collected. A variety of measures were completed including a knowledge monitoring assessment (KMA) metacognition measure (Valkanova & Watts, 2007). The researchers created the KMA measure used in the study with the English language STEM vocabulary being taught at the summer camp and additional English language academic vocabulary. The KMA consisted of 12 subscales and total scores measuring the children's ability to read the words, define the words, and to create sample sentences for 3 lists (i.e., STEM vocabulary easy, STEM vocabulary harder, and Dolch sight words). For this demonstration study, the following KMA variables were used: reading STEM vocabulary easy (Oral1), reading STEM vocabulary harder (Oral2), defining STEM vocabulary easy (Def1), harder STEM words (Def2), Dolch words (Def3), creating sample sentences with easy STEM words (Samp1), and harder STEM words (Samp2), and with Dolch words (Samp3).

The Classroom Language Interaction Checklist, Third Edition (CLIC-3; Collier, 2010) was used to measure participants' language abilities in English and Spanish in the classroom. The CLIC-3 is composed of 10 items to assess social language skills and 55 items to assess overall academic language skills. The participants' summer camp teachers completed the CLIC-3 rating forms, which the examiners then used to determine an estimated cognitive academic language proficiency (CALP) score (scores can range from 1 to 6) for English language use. For the purposes of this demonstration study, scores measuring English CALP (CALPEng) use of social language in English (SocLangEng) and use of academic language in English (AcaLangEng) were utilized. The participants were provided instruction in English-only during the summer camp and the KMA is an English-only assessment in this context. Systematic observations were conducted using the School Psychology Tools application for iPad. Trained raters observed each individual for 15 minutes and used a partial interval recording method to record the frequency of the following behaviors: passive teacher engagement (e.g., student following the teacher's instructions; Pas_Eng) and active teacher engagement (e.g., student interacting directly with the teacher; Act_Eng).

The outcome variable of interest in this study was the total score on the Kaufman Brief Intelligence Battery, Second Edition (KBIT-II; Kaufman & Kaufman, 2004). The KBIT-II provided an estimate of participants' cognitive ability using both verbal and nonverbal estimates of intelligence. Participants completed tasks that measured receptive and expressive language, verbal comprehension, reasoning, and novel problem solving skills. The goal of the analysis for this specific dataset was to identify factors from among the KMA, CLIC-3 subscales, as well as the observation scores that significantly predict the total intelligence score for the migrant worker children. Easily observable and recorded variables that are good predictors of cognitive ability may be helpful for quickly screening children who may be on either end of the exceptionality continuum and; thus, may be in need of further assessment, classroom teaching differentiation, or targeted methods of instruction and/or acceleration. For this data, there were a total of 13 independent variables, matching the size of the 13 participants, and analyses were conducted using OLS regression as well as PCR, SPCR, and PLSR. The broader goal of the study, as mentioned previously, was to demonstrate the utility of three methods for conducting regression for high dimensional data and to compare the results obtained by these methods.

Results

The means and standard deviations of the raw scores for all variables used in the analyses appear in Table 1. Prior to application of the regression analyses, all variables used in the analysis were standardized. The Pearson's *r* correlation coefficients between each of the independent variables and the KBIT-II (Kaufman & Kaufman, 2004) score appear in the final column of Table 2. From these, we can conclude that scores on active engagement, KMA: Samp 2; KMA: Def 3; and KMA: Samp 3, were all positively correlated with the KBIT-II above 0.3, or in Cohen's moderate correlation range (Cohen, 1988). In addition, KMA: Oral 2 was also correlated with the KBIT-II in the moderate range, though, in the negative direction. Of the remaining variables, only KMA: Oral 1 was negatively associated with KBIT-II, whereas the others had positive correlations in what Cohen termed the small range.

Table 1. *Means and Standard Deviation (SD) of Variables used in the Analysis for Migrant Children*

Variable	Mean	SD
Active Engagement	28.7	25.7
Passive Engagement	24.0	35.1
CALP	4.8	0.8
Social Language	12.1	2.8
Academic Language	34.3	11.6
KMA: Oral 1	9.8	0.5
KMA: Oral 2	8.9	1.1
KMA: Def 1	10.0	4.3
KMA: Def 2	7.5	3.1
KMA: Def 3	12.8	3.1
KMA: Samp 1	13.3	3.5
KMA: Samp 2	7.9	4.0
KMA: Samp 3	16.4	2.6
KBIT-II	91.5	17.2

Note: *(n=13)*

OLS

As mentioned above, the goal of this study was to estimate a prediction model for the KBIT-II. Thus, the first approach used was multiple regression, which might be considered the default analysis strategy for this research problem. Table 2 includes results from a standard OLS regression in which the KBIT-II served as the response variable.

All variables were entered in one step. First, it should be noted that standard errors could not be estimated for the model coefficients because the sample size was equal to the number of independent variables. Second, there is not a coefficient estimate for KMA: Samp 3, the last independent variable in

Table 2. *Coefficients for Standardized Independent Variables*

Variable	OLS (R^2=NA)	PCR (R^2=0.23)[a]	SPCR (R^2=0.40)[a]	PLSR (R^2=0.51)[a]	Correlation with KBIT
Active Engagement	1.11	-0.01	0.66	0.36	0.38
Passive Engagement	3.35	0.01	-[b]	0.11	0.27
CALP	0.90	0.04	-	-0.07	0.13
Social Language	-5.95	0.04	-	-0.08	0.23
Academic Language	4.02	0.04	-	-0.02	0.19
KMA: Oral 1	2.05	-0.08	-	0.02	-0.20
KMA: Oral 2	-3.19	-0.06	-0.60	-0.37	-0.35
KMA: Def 1	-3.61	0.08	-	-0.29	0.12
KMA: Def 2	5.12	0.05	-	0.12	0.23
KMA: Def 3	2.78	0.10	0.56	-0.01	0.32
KMA: Samp 1	-2.36	0.03	-	0.12	0.14
KMA: Samp 2	-4.61	0.03	0.72	0.36	0.41
KMA: Samp 3	NA	0.07	0.52	0.37	0.30
Intercept	0.00	-0.00	-0.00	-0.00	
Component 1		-0.16 (0.15)	0.42 (0.35)	0.31 (0.13)*	
Component 2		0.01 (0.17)	0.73 (0.35)*	0.26 (0.18)	
Component 3		0.10 (0.22)	0.45 (0.52)	0.40 (0.31)	

Note. *$p < 0.05$; [a]For PCR, SPCR, and PLSR three components were retained; [b]Variables not retained in the reduced SPCR model do not have a regression coefficient.

the list, again because the number of observations equaled the number of independent variables. Finally, because the model could not be properly estimated, there was not an estimate for the total R^2. With regard to the coefficients themselves, which are standardized beta weights, the values are generally quite large and in some cases inconsistent with the signs of the correlation coefficients. Taken together, these results indicate that the use of OLS in this instance was clearly not appropriate. Therefore, an alternative approach is necessary.

PCR

The first data reduction analysis to be used was PCR. Based on the results of PA, as well as the proportion of variance extracted, we concluded that 3 components should be retained. Taken together, these accounted for approximately 73% of the variance in the *x* variables as can be seen at the bottom of Table 3.

The first component accounted for 34% of the variance, followed by the second at 24%, and the third with 15%. The component loadings for each *x* with each component appear in Table 3. Based on recommendations in the literature (e.g., Tabachnick & Fidell, 2013), loading absolute values greater than or equal to 0.32 were taken to show that a variable was associated with a component in a meaningful way. Based on this criterion, Component 1 was primarily associated with the KMA: Def and KMA: Samp variables, while Component 2 was associated with the CLIC-3 CALP, Social language, and Academic Language scores. Finally, Component 3 was primarily associated with the KMA: Oral scores (STEM word reading). An examination of the loadings included in Table 3 reveals that none of the predictor variables exhibited cross loadings greater than the 0.32 cut value (Tabachnick & Fidell). Thus, each predictor can be said to load on to only one of the components, based on this criterion.

Subsequent to the identification of the number and nature of the principal components, these values were then used as independent variables in a regression analysis in which the KBIT-II comprehensive score served as the dependent variable. The results, including coefficient estimates for each component as well as their standard errors, appear in the third column and final 3 rows of Table 2. None of the components were significantly related to the KBIT-II score. Thus, based on these results, we would conclude that there are not significant relationships between the various independent measures and overall intelligence as measured by the KBIT-II. Together, the 3 components accounted for approximately 23% of the variance in the KBIT-II (R^2=0.23). Finally, (6) was applied to these results in order to obtain PCR coefficient estimates for the individual *x*s. The relatively small magnitude of these coefficients further reinforces the finding that there are not strong relationships between the set of independent variables and the dependent, at least in the context of PCR. However, given the presence of the moderate relationships that were identified with the correlation coefficients, this result may be called into question.

Table 3. *Rotated PCR Component Loadings: 3 Component Solution*

Variable	Component 1	Component 2	Component 3
Active Engagement	0.04	0.03	-0.03
Passive Engagement	-0.24	-0.32	-0.11
CALP	-0.22	0.42	-0.29
Social Language	-0.15	0.45	-0.21
Academic Language	-0.23	0.43	-0.31
KMA: Oral 1	0.03	-0.24	-0.57
KMA: Oral 2	-0.06	-0.31	-0.51
KMA: Def 1	-0.37	-0.09	0.31
KMA: Def 2	-0.37	-0.20	0.05
KMA: Def 3	-0.41	0.15	0.26
KMA: Samp 1	-0.30	-0.25	0.00
KMA: Samp 2	-0.35	-0.19	-0.10
KMA: Samp 3	-0.41	0.05	0.00
Proportion variance	0.34	0.24	0.15
Cumulative proportion variance	0.34	0.58	0.73

The first component accounted for 34% of the variance, followed by the second at 24%, and the third with 15%. The component loadings for each *x* with each component appear in Table 3. Based on recommendations in the literature (e.g., Tabachnick & Fidell, 2013), loading absolute values greater than or equal to 0.32 were taken to show that a variable was associated with a component in a meaningful way. Based on this criterion, Component 1 was primarily associated with the KMA: Def and KMA: Samp variables, while Component 2 was associated with the CLIC-3 CALP, Social language, and Academic Language scores. Finally, Component 3 was primarily associated with the KMA: Oral scores (STEM word reading). An examination of the loadings included in Table 3 reveals that none of the predictor variables exhibited cross loadings greater than the 0.32 cut value (Tabachnick & Fidell). Thus, each predictor can be said to load on to only one of the components, based on this criterion.

Subsequent to the identification of the number and nature of the principal components, these values were then used as independent variables in a regression analysis in which the KBIT-II comprehensive score served as the dependent variable. The results, including coefficient estimates for each component as well as their standard errors, appear in the third column and final 3 rows of Table 2. None of the components were significantly related to the KBIT-II score. Thus, based on these results, we would conclude that there are not significant relationships between the various independent measures and overall intelligence as measured by the KBIT-II. Together, the 3 components accounted for approximately 23% of the variance in the KBIT-II (R^2=0.23). Finally, (6) was applied to these results in order to obtain PCR coefficient estimates for the individual *x*s. The relatively small magnitude of these coefficients further reinforces the finding that there are not strong relationships between the set of independent variables and the dependent, at least in the context of PCR. However, given the presence of the moderate relationships that were identified with the correlation coefficients, this result may be called into question.

SPCR

As noted above, a potential weakness of the PCR approach is that it does not take into account relationships between individual *x* variables and *y* when forming the principal components. As a result, while these components may well account for a sizable share of the variance in the independent variables, as is the case here, they may have relatively little relationship with the dependent variable, again as is the case in this example. Therefore, an alternative approach would be SPCR in which *x*s that are related to *y* in the context of simple linear regression are first identified and then subsequently included in a PCA in order to reduce the overall dimensionality in the data, as described above. The first step in conducting SPCR is to determine the appropriate threshold (*t*) value for a variable's inclusion in the analysis. As noted above, this is done using the jackknife procedure to find the optimal *t*. Figure 1 displays values of *t* on the *x*-axis and the Likelihood ratio statistic on the *y*-axis. From this graph, it appears that 0.5 is the optimal threshold value and was, therefore, used throughout the SPCR analysis.

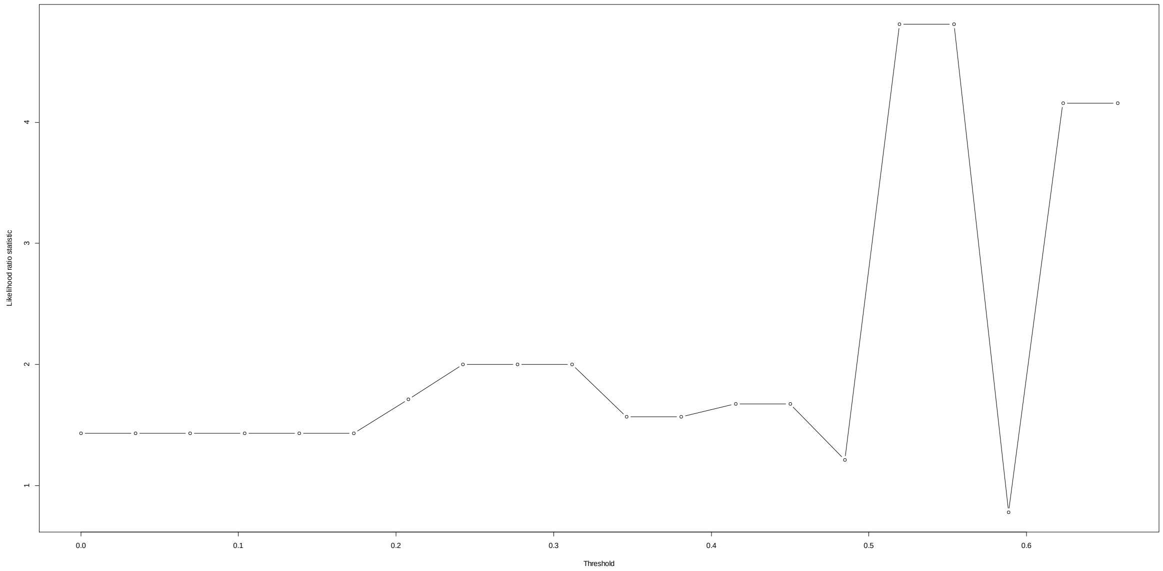

Figure 1. Likelihood ratio statistic by threshold value.

Three components were retained for estimating the regression model with the KBIT-II comprehensive score as the dependent variable. The results of this analysis appear in the last three rows and fourth column of Table 2. Component 2 was statistically significantly related to the KBIT-II score, with a positive coefficient, indicating that higher Component 2 scores were associated with higher scores on the KBIT-II. More specifically, a 1 unit increase in the Component 2 score was associated with a 0.73 unit increase in the standardized KBIT-II, or approximately 0.73 of a standard deviation. The three components together accounted for approximately 40% ($R^2 = 0.40$) of the variance in the KBIT-II score.

As with the results from the PCR, it is possible to obtain coefficient estimates for the individual independent variables, based on the coefficients for the components, as well as the eigenvectors linking the components to the *x*s. As noted above, in SPCR only, those variables whose relationships with y are sufficiently large will be included in the principal components analysis. Therefore, it is only for these variables that coefficient estimates can be obtained. An examination of Table 2, column 4 reveals the coefficient values for the variables that were included in the final reduced SPCR model. Five of the original *x* variables had relationships with y that exceeded the threshold of 0.5 and were, thus, included in the model: Active engagement, KMA: Oral 2, KMA: Def 3, KMA: Samp2, and KMA: Samp 3. These were also the variables having the highest correlation coefficients with the KBIT-II, as described previously. The coefficient estimates in Table 2 are □ weights. Of these variables, Active Engagement, Def 3, Samp 2, and Samp 3 were positively associated with the KBIT-II comprehensive score, whereas Oral 2 had a negative relationship, just as was seen with the Pearson's *r* values. Additionally, the SPCR coefficients were much larger in value than those from PCR, for the same variables. Indeed, the coefficients associated with these five variables were larger than any of those estimated using PCR.

In addition to obtaining estimates for the relationships between the components and *y*, and estimates of the relationships between the individual *x* variables and y given the components, we may also wish to know the relative importance of each retained independent variable with respect to the individual components. As described above, this can be calculated using SPCR variable importance measures, where relatively larger values indicate that the variable was more important in terms of forming the component. It must be understood that these values are not component loadings akin to those we discussed in the context of PCR, but they do serve a very similar role in that they provide insights regarding the nature of each component. The importance measures for this analysis appear in Table 4. From these, we can see that Samp 3, Def 3, and Samp 2 were the most important variables in the formation of Component 1. On the other hand, Oral 2 and Active Engagement were the most important for Component 2, with Samp 2 and Samp 3 playing very little role in the formation of this second component. Finally, Component 3 was most associated with Samp 2 and Active Engagement, followed by Oral 2.

Revisiting the regression results described above, we can conclude that the component associated with Oral 2 and Active Engagement represented the one statistically significant relationship that was found in the data. This did not mean, however, that the coefficient estimates for these two variables were necessarily the largest among the retained *x*s, though, they were not the smallest either.

Table 4. *Variable Importance Measure for SPCR*

Variable	Component 1	Component 2	Component 3
Active Engagement	-2.27	5.55	3.72
Passive Engagement	0.00	0.00	0.00
CALP	0.00	0.00	0.00
Social Language	0.00	0.00	0.00
Academic Language	0.00	0.00	0.00
KMA: Oral 1	0.00	0.00	0.00
KMA: Oral 2	1.54	-7.00	2.42
KMA: Def 1	0.00	0.00	0.00
KMA: Def 2	0.00	0.00	0.00
KMA: Def 3	7.75	3.37	-1.33
KMA: Samp 1	0.00	0.00	0.00
KMA: Samp 2	6.00	0.11	3.88
KMA: Samp 3	8.46	-0.40	-0.98

PLSR

As with SPCR, PLSR is a methodology that seeks to reduce the number of predictors in a regression equation by creating linear combinations of the original set of *x* variables while also accounting for their relationships with *y*. However, unlike SPCR, PLSR does not reduce the number of candidate *x*s that might be used in forming these linear combinations, but rather uses all of the original independent variables. The first step in PLSR is to determine the number of components to retain. An examination of Figure 2 reveals that the RMSEP (i.e., prediction error) decreases in a fairly linear manner as the number of components increases.

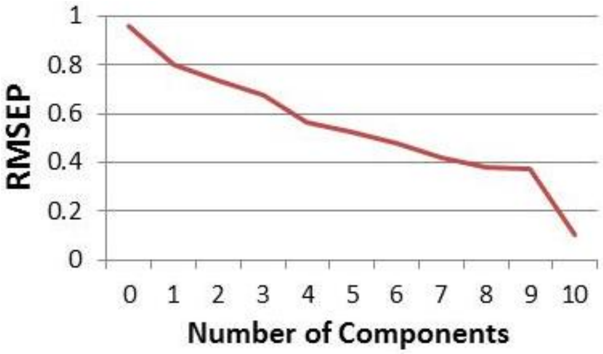

Figure 2. RMSEP by number of components.

The proportion of variance explained in the independent and dependent variables by each component appears in the final 2 rows of Table 5. The first component accounted for 29% of the variance in the *x* variables, and 30% in *y*, while component 2 accounted for an additional 17% in the *x*s, and 12% in *y*, and a third component accounted for 16% additional variance in the independent variables, and 9% in *y*. Given that the 4th component only accounted for 7% of the variance, we decided to retain 3 components.

The component loadings for PLSR appear in Table 5. As with the PCR component loadings in Table 3, these values provide information regarding which of the *x* variables were most associated with each of the components. We will use the same 0.32 criteria for identifying variables that are most related to each component. Component 1 was most strongly associated with Active engagement, KMA: Oral 2, KMA: Def 3 and KMA: Samp2. Component 2 was most strongly associated with Active Engagement, KMA: Oral 2 and KMA: Def 1. Component 3 was most strongly associated with Social Language, KMA: Oral 1 and KMA: Samp 3. Two of the original predictors, Active Engagement and KMA: Oral 2, exhibited cross loadings above the 0.32 cut value based on the results in Table 5.

The regression results for the three retained components appear in the last three rows and fifth column of Table 2. Only the first PLSR component was significantly related to the KBIT-II score, with a positive coefficient of 0.31. This result indicates that for every 1 unit in the standardized Component 1 value, there would be a 0.31 increase in the standardized KBIT-II score. Together, the three PLSR components accounted for approximately 51% of the variance in the KBIT-II (R^2 = 0.51). The standardized coefficients for each independent variable, given the PLSR solution, appear in Table 1. Since Component 1 was the only linear combination that had a significant relationship with the KBIT-II, it is not surprising that several variables most associated with this component also had the largest coefficient values, including Active Engagement, KMA: Oral 2, and KMA: Samp 2. In addition, KMA: Samp 3 and KMA: Def 1 also had relatively large coefficients.

Table 5. *PLSR Component Loadings: 2 Component Solution*, *Values less than 0.1 are coded as -

Variable	Component 1	Component 2	Component 3
Active Engagement	0.39	0.48	-0.17
Passive Engagement	0.28	-*	-
CALP	0.14	-0.27	-
Social Language	0.24	-0.11	-0.39
Academic Language	0.20	-0.20	-
KMA: Oral 1	-0.21	-	0.38
KMA: Oral 2	-0.36	-0.44	-
KMA: Def 1	0.12	-0.49	-0.31
KMA: Def 2	0.23	-0.21	0.23
KMA: Def 3	0.34	-0.30	-0.12
KMA: Samp 1	0.15	-0.22	-
KMA: Samp 2	0.42	-	0.27
KMA: Samp 3	0.31	-0.13	0.65
Proportion variance in *x*	0.29	0.17	0.16
Proportion variance in *y*	0.30	0.12	0.09

Finally, the Pearson's correlation coefficients among the component scores for PCR, PLS, and SPCR appear in Table 6. From these we can see that the strongest relationships were between scores derived from PCR and PLS, with positive correlations among components 2 and 3 and negative correlations for component 1. A similar pattern of correlations was present between component scores from SPCR and PCR, though, the values were somewhat smaller in magnitude. The relationships between the 3 primary components from SPCR and PLS were positive and ranged from 0.57 to 0.65.

Table 6. *Pearson Correlation Coefficients among Component Scores for PCR, PLS, and SPCR*

		PLS	
PCR	Component 1	Component 2	Component 3
Component 1	-0.76	0.29	-0.03
Component 2	0.28	0.71	-0.30
Component 3	0.06	0.29	0.72
		SPCR	
PCR	Component 1	Component 2	Component 3
Component 1	-0.58	0.42	-0.21
Component 2	0.41	0.65	0.31
Component 3	0.02	-0.02	0.82
		PLS	
SPCR	Component 1	Component 2	Component 3
Component 1	0.65	-0.30	0.25
Component 2	-0.34	0.57	0.14
Component 3	-0.04	0.04	0.59

Discussion

High dimensional data presents the researcher with some unique challenges. In particular, when the number of independent variables is nearly as large as, or larger than the sample size, problems of overfitting and high parameter estimate variance may be encountered. Indeed, if the number of independent variables exceeds the sample size, it will not be possible to obtain a full set of regression coefficient estimates and associated standard errors. Nonetheless, there are situations in which such high dimensional data are all that is available, so the researcher must find an approach for analyzing them. The goal of this study was to describe and demonstrate in some detail three alternatives for dimension reduction and regression analysis that could be used in the presence of high dimensional data. The first of these, PCR, is relatively well known, but also has some fairly significant limitations. In particular, it maximizes variance accounted for by the principal components in the *x* variables only, so it may not yield components that are optimally related to *y*. As a consequence, two other approaches were discussed here, SPCR and PLSR, which may prove to be more appropriate for the high dimensional data situation because they identify components of the *x*s taking into account their relationships with *y*. In the case of SPCR, this is done by first identifying variables that are associated with *y* and then performing principal components analysis only on them. On the other hand, PLSR creates linear combinations of the independent variables by maximizing not only the variance explained in them, but also the covariances of the independent variables with the dependent. In both cases, the regression equations involving the linear combinations should yield stronger relationships with *y* than is the case for PCR (Yu et al., 2006). The results presented above would appear to support this proposition, as PCR did not yield any significant relationships between the components and the dependent variable, whereas SPCR and PLSR each did.

In terms of comparing and contrasting results for SPCR and PLSR, several points should be raised. First, each had a single component that was significantly related with scores on the KBIT-II. Based on the importance measures for SPCR, this component was primarily associated with KMA: Oral 2 and Active Engagement, with KMA: Def 3 playing a somewhat less important, but still meaningful role in determining the component's value. For PLSR, the significant component was also primarily associated with Active Engagement KMA: Oral 2 and KMA: Def 3. In addition, however, this component was also associated with KMA: Samp 2, which did not play a large role in the significant component in SPCR. In terms of the specific coefficients for the individual *x* variables, there was some overlap in terms of which variables each identified as being most strongly associated with the KBIT-II. Both SPCR and PLSR found Active Engagement, KMA: Oral 2, KMA: Samp 2, and KMA: Samp 3 to have relatively large standardized coefficient values. On the other hand, SPCR also identified KMA: Def 3 as being fairly strongly associated with the KBIT-II, whereas PLSR did not. Conversely, PLSR yielded a relatively large value for KMA: Def1 and SPCR did not. With respect to the nature of these relationships, results from both methods indicated that higher levels of Active Engagement were associated with higher KBIT-II intelligence test scores as were higher scores on KMA: Samp 2 and KMA: Samp 3. In addition, both methods found that higher scores on KMA: Oral 2 were associated with lower scores on the KBIT-II. Finally, SPCR results showed that individuals with higher KMA: Def 3 scores had higher KBIT-II scores, and PLSR found that those with higher KMA: Def 1 scores had lower values for the KBIT-II. It should also be noted that the standardized coefficients for PCR were smaller than those of the other methods, which is in keeping with the lack of significant relationships between the KBIT-II and the PCR components. With regard to OLS, the standardized coefficients generally took extreme values such that their interpretation would be difficult at best.

Limitations and Directions for Future Research

As with all research, there are limitations to the current study that must be acknowledged and directions for future research that can extend this work. First of all, because this example utilized an existing data set, and not a simulation, it is not possible to know the actual population parameters. Thus, while SPCR and PLSR both indicated that multiple predictor variables were in fact related to the outcome, whereas PCR did not, we cannot know for certain which result is correct. It is possible that the stronger relationships between some predictors and the dependent variable found by the two alternatives to PCR were in fact capitalizing on chance. That is, they both strive to first identify significant predictors before reducing the data dimensionality into a small number of components and that PCR was correct in finding no significant regression results. Thus, future research should expand upon this work both by using Monte Carlo simulation techniques, as well as by analyzing additional high dimensional datasets, in order to better understand the comparative performance of these methods. While SPCR and PLSR do have a number of apparent advantages over PCR, their use and interpretation of results must be conducted with caution. In particular, although both methods offer the important advantage of being able to estimate regression models for research scenarios in which other approaches are unable to do so, even these methods can suffer estimation problems when the number of variables far exceeds the number of observations. In such instances, the correlation matrix among the predictors will not be positive semi-definite and even SPCR and PLSR may be unable to converge on a solution. Thus, while these approaches do provide a definite advantage in many instances involving high dimensional data, they are not impervious to problems associated with the p>>n case.

Conclusions and Recommendations for Practice

The goal of this research was to demonstrate methods for conducting regression when the number of independent variables is nearly as large as, or larger than, the number of observations. Traditionally, researchers have either broken such a problem down into a series of univariate regression models, or employed principal components analysis to reduce the number of independent variables and then applied regression to those. However, we have presented two alternatives that might prove more useful than either of these methods. The analysis of the migrant worker children data revealed the relative ease with which both SPCR and PLSR can be employed and the types of results that the researcher can expect to obtain from them. Both SPCR and PLSR yielded statistically significant relationships between the components and the dependent variable, whereas PCR did not. This result is consistent with what might be expected, given that both approaches take into consideration the relationships of the *x*s with *y* when creating the linear combinations, while PCR does not. SPCR and PLSR can be carried out using the R software package, with relative ease, and the R code necessary to conduct the analyses demonstrated here appear in the appendix to this paper.

In considering the results presented here, the researcher faced with high dimensional data should keep in mind the following issues when deciding which approach might be optimal for a given research situation. First, among these methods, SPCR and PLSR each require more decisions be made during data analysis than does PCR. For example, with SPCR, the researcher must use a threshold value to determine which predictors will be included in the analysis. Thus, threshold selection is a crucial step in applying this technique as it determines the number of predictors that will be included in the final model. As noted above, the decision regarding the optimal value of this threshold is based upon the results of a jackknife cross-validation procedure. However, when the sample is extremely small, this cross-validation technique could be somewhat unstable, making the determination of the optimal threshold difficult. Similarly, the use of jackknife cross-validation in determining the number of components to retain with PLSR may also be unstable with very small samples. In addition, unlike PCR, SPCR purposefully excludes many, if not most, of the original set of independent variables from its final analysis. Therefore, if the researcher is interested in obtaining parameter estimates for each independent variable, SPCR may not be the optimal methodology. Rather, PLSR or PCR may be more appropriate. Another factor in the decision regarding which method to use is the extent to which the researcher wants to maximize understanding of the interrelationships among the independent and dependent variables, versus the extent to which she would like to first gain insights into the structure underlying the independent variables, before ascertaining their relationship(s) with the outcome of interest. If the primary goal is in identifying the optimal predictive model, then PLSR or SPCR may offer the better path because they are explicitly designed to maximize the amount of shared variance explained by the model. On the other hand, if the researcher first wants to understand what structure underlies the predictors prior to determining how they might be related to the outcome, then PCR may be optimal. Finally, interpretation of the cross-validation results for both SPCR and PLSR is subjective in nature, so coming to a definitive conclusion regarding the optimal solution will inherently carry with it some lack of precision. On the other hand, as was noted previously, PCR may have difficulty finding parameter estimates when the sample size is very small due to a non-positive, semi-definite correlation matrix. However, SPCR, in particular, may be less likely to suffer from this problem because it reduces the number of predictors to be considered through the initial univariate analyses that are a part of its algorithmic structure.

In summary, each method included in this study offers the interested researcher something useful when dealing with high dimensional data. While it is difficult to provide broad, sweeping recommendations for practice based on a single analysis, we do offer the following general advice to researchers faced with high dimensional data. If the number of predictors is much, much larger than the sample size, SPCR may offer the optimal analysis solution because with judicious selection of the threshold, it will reduce the number of predictors prior to fitting any regression models. Thus, making SPCR more likely to avoid a non-positive definite correlation matrix. If the researcher is faced with high dimensional data, but *p* is not extremely larger than *n*, and the goal of the study is to optimize the prediction model, then PLSR may be the best choice, as it does not automatically remove a large number of predictors (as SPCR typically does). Though, like SPCR, PLSR does reduce the number of predictors included in the final model, while accounting for their relationships with the dependent variable. Finally, if the researcher's primary goal is to understand the underlying structure among the predictors, and only after doing so to fit a regression model, then PCR may be the optimal choice for analysis.

References

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association, 101,* 119-137.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory, and applications.* New York: Springer.

Carin, L., Hero, A., Lucas, J., Dunson, D., Minhua, C., Henao, R., Tibau-Piug, A., Zaas, A., Woods, C.W., & Ginsburg, G.S. (2012). High-dimensional longitudinal genomic data: An analysis used for monitoring viral infections. *Signal Processing Magazine, IEEE, 29,* 108-123.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collier, C. (2010). *Classroom language interaction checklist* (3rd ed.). Lake Ferndale, WA: Cross Cultural Developmental Education Services.

Datta, S. (2001). Exploring relationships in gene expressions: A partial least squares approach. *Gene Expression, 9,* 257-264.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: John Wiley & Sons.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics tools (with discussion). *Technometrics, 35*, 109-148.

Gupta, R., & Kabundi, A. (2010). Forecasting real U.S. house prices: Principal components versus Bayesian regressions. *International Business & Economics Research Journal, 9, 141-152.*

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning.* New York: Springer.

Horn J. L. (1965). A rationale and a test for the number of factors in factor analysis. *Psychometrika*, *30* 179–185.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417-441.

Kaufman, A. S., & Kaufmann, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.).Bloomington, MN: Pearson.

Moon, H., Ahn, H., Kodell, R.L., Baek, S., Lin, C.J., & Chen, J.J. (2007). Ensemble methods for classification of patients for personalized medicine with high-dimensional data. *Artificial Intelligence in Medicine, 41,* 197-207.

Ongel, A., Kohler, E., & Harvey, J. (2008). Principal components regression of onboard sound intensity levels. *Journal of Transportation Engineering, 134,* 459-466.

Rosipal, R., & Trejo, L.J. (2001). Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research, 2,* 97-123.

Samet, H. (2006). *Foundations of multidimensional and metric data structures.* San Francisco: Morgan-Kaufmann.

Solomon, A. C., Hill, R, Jassen, E., Sanders, S. A., & Hieman, J. R. (2012, January). *Uniqueness and how it impacts privacy in health-related social science datasets*. Paper presented at IHI, Miami.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Boston: Pearson.

Tan, Y., Shi, L., Tong, W., & Wang, C. (2005). Multi-class cancer classification by Total Principal Component Regression (TPCR) using microarray gene expression data. *Nucleic Acids Research, 33,* 56-65.

United States Department of Justice & Department of Education. Civil Rights Division, Office for Civil Rights & Office of the General Counsel. (2011). *Joint 'dear colleague' letter*. Retrieved from http://www2.ed.gov/about/offices/list/ocr/letters/colleague-201101.html

Valkanova, Y., & Watts, M (2007). Digital story telling in a science classroom: Reflective self-learning in action. *Early Child Development and Care, 177*, 739-807.

Weiss, A., Gale, C.R., Batty, G.D., & Deary, I.J. (2013). A questionnaire-wide association study of personality and mortality: The Vietnam experience study. *Journal of Psychosomatic Research, 74,* 523-529.

Yu, S., Yu, K., Tresp, V., Kriegel, H-P., & Wu, M. (2006). Supervised probabilistic principal components analysis. *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 464-473.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu